Capstone Project

Introduction

This project is going to use Jupyter Notebook to analyze GEO data of a Sydney suburb named Beecroft.

As part of the Data Science assignment, it'll use public available data as well as various libraries and tools. The Foursquare location data will be used during the exercise.

Business Problem

Often people feel lost when trying to make decision based on location. The data is either not available or not well organized. This project is trying to resolve this problem by analysing various related data.

Based on the detail analysis, interesting party may make business decision with more information and confidence. For example, a potential property investor may purchase a property in suburbs which have more desired facilities. A small business may open a coffee shop in a suburb to maximize the profit.

The target audience for the report are home buyers and business investors.

Data Description

The Sydney metro data used is from this link:

https://www.prospectshop.com.au/Files/SydneyMetro_Postcodes.xls

The sample data:

	id	postcode	suburb	state	latitude	longitude
0	398	1001	Sydney	NSW	-33.79	151.27
1	399	1002	Sydney	NSW	-33.79	151.27
2	400	1003	Sydney	NSW	-33.79	151.27
3	401	1004	Sydney	NSW	-33.79	151.27
4	402	1005	Sydney	NSW	-33.79	151.27

The original data format will be transformed to work together with the other APIs. The Data filtering phase:

- Format the column headers
- Remove duplicate suburbs since one suburb may have more entries of postcodes due to the post box.

- Remove suburb without GEO data
- Remove PO box entries which have postcode start with 1000

The location and venue data come from the Foursquare.

Postcode data:

https://gist.github.com/randomecho/5020859

Need to import the data from the above URL into MySQL and then export the data as .csv file. This involves to install MySQL in a Ubuntu VM, execute the SQL statement from the above link, export data as .csv.

Methodology

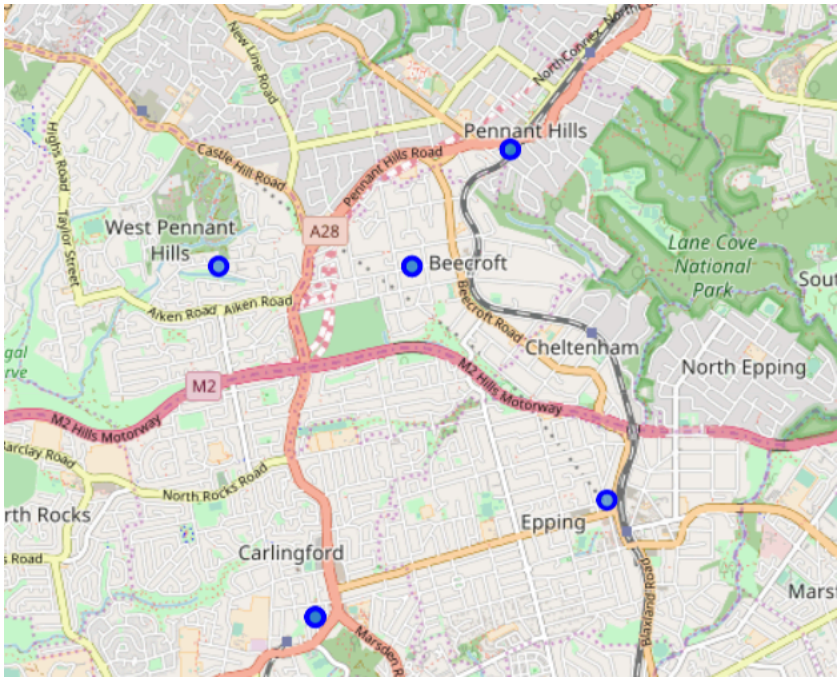
The focus of the report is to analyse the neighbourhood of Beecroft including all interesting venues data obtained from the FourPoints website. The venues data were sorted and grouped by their categories. The output table should give some ideas to the readers what kind of lifestyle is in each suburb. For example, some suburbs have more Café while others may have more grocery shops, some suburbs have more parks while others may have more bus stops.

The sample output:

	PostCode	Neighborhood	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	2076	Normanhurst	151.10	-33.72	1	Café	Gym	Indian Restaurant	Asian Restaurant	Park	Chinese Restaurant
1	2076	North Wahroonga	151.10	-33.72	0	Café	Gym	Indian Restaurant	Asian Restaurant	Park	Chinese Restaurant
2	2076	Wahroonga	151.10	-33.72	0	Café	Gym	Indian Restaurant	Asian Restaurant	Park	Chinese Restaurant
3	2117	Dundas	151.04	-33.80	0	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant
4	2117	Dundas Valley	151.04	-33.80	1	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant
5	2117	Oatlands	151.04	-33.80	0	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant
6	2117	Telopea	151.04	-33.80	0	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant
7	2118	Carlingford	151.05	-33.78	0	Fast Food Restaurant	Café	Train Station	Supermarket	Italian Restaurant	Shopping Mall
8	2118	Carlingford Court	151.05	-33.78	1	Fast Food Restaurant	Café	Train Station	Supermarket	Italian Restaurant	Shopping Mall
9	2118	Carlingford North	151.05	-33.78	0	Fast Food Restaurant	Café	Train Station	Supermarket	Italian Restaurant	Shopping Mall

The detail steps and report can be found from the Jupyter Notebook.

The Folium map is used to visualize the GEO locations. For example, the main target suburb and its neighbourhood:



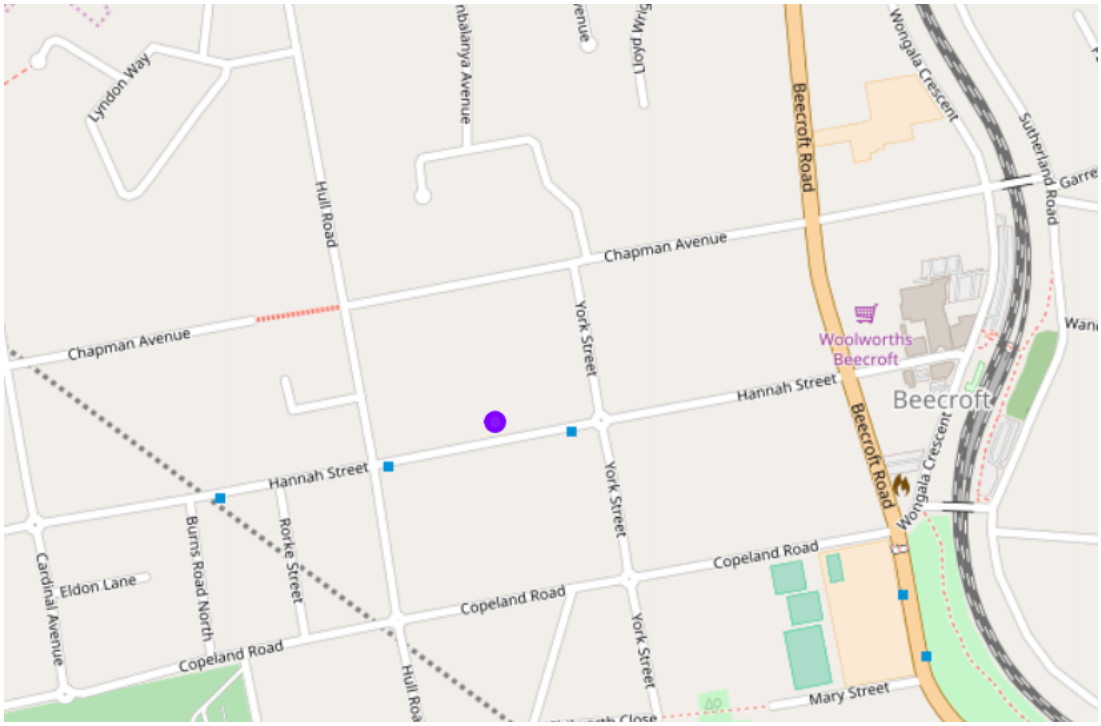
Machine learning technics

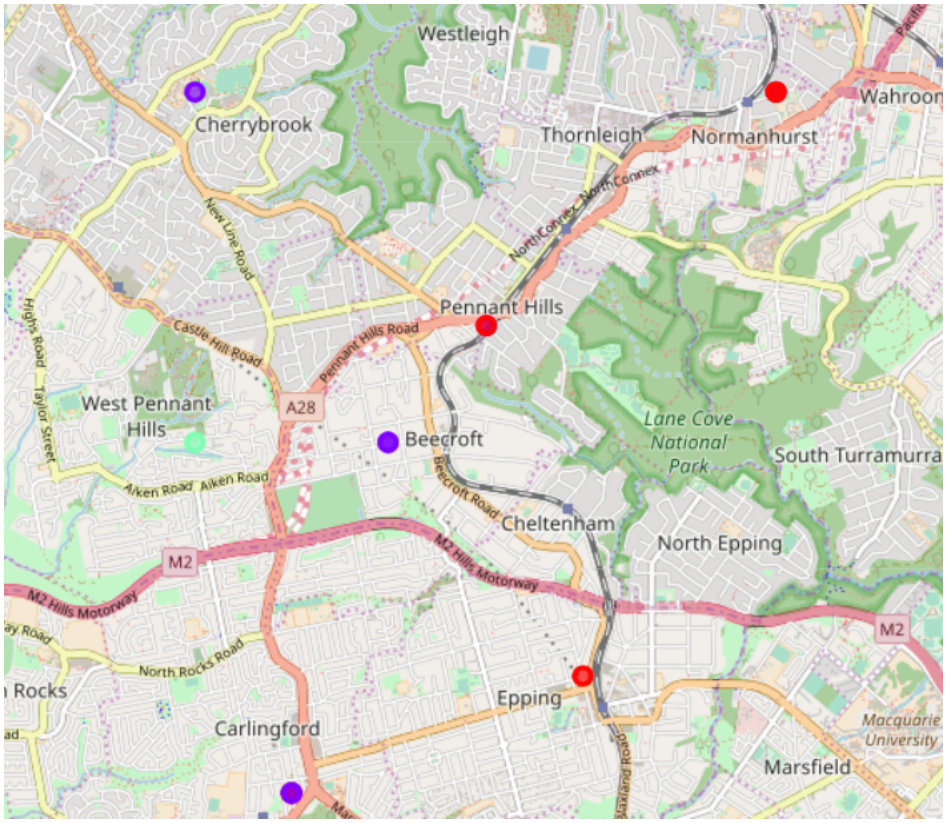
Used *k*-means to cluster the neighborhood into 3 clusters. We can group the suburbs according to their similarities in terms of the categories of their venues. I've added more nearby suburbs due to the small quantity of data.

Results

The final results are presented in the Notebook. It includes a map and a few tables of K-means cluster for the neighbourhood.

Sample map:





Sample K-mean cluster:

Cluster 1

```
eecroft_merged['Cluster Labels'] == 0, beecroft_merged.columns[[1] + list(range(5, beecroft_merged.shape[1]))
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	North Wahroonga	Café	Gym	Indian Restaurant	Asian Restaurant	Park	Chinese Restaurant	Bakery	Grocery Store	Electronics Store	Japanese Restaurant
2	Wahroonga	Café	Gym	Indian Restaurant	Asian Restaurant	Park	Chinese Restaurant	Bakery	Grocery Store	Electronics Store	Japanese Restaurant
3	Dundas	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant	Sandwich Place	Chinese Restaurant	Soccer Field	Bus Station
5	Oatlands	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant	Sandwich Place	Chinese Restaurant	Soccer Field	Bus Station
6	Telopea	Pizza Place	Grocery Store	Café	Supermarket	Pub	Fast Food Restaurant	Sandwich Place	Chinese Restaurant	Soccer Field	Bus Station
7	Carlingford	Fast Food Restaurant	Café	Train Station	Supermarket	Italian Restaurant	Shopping Mall	Gym	Gym / Fitness Center	Food Court	Korean Restaurant
9	Carlingford North	Fast Food Restaurant	Café	Train Station	Supermarket	Italian Restaurant	Shopping Mall	Gym	Gym / Fitness Center	Food Court	Korean Restaurant
14	Westleigh	Café	Pizza Place	Thai Restaurant	Liquor Store	Chinese Restaurant	Platform	Gym / Fitness Center	Fast Food Restaurant	Shopping Mall	Bakery
15	North Epping	Thai Restaurant	Indian Restaurant	Pizza Place	Italian Restaurant	Social Club	Fried Chicken Joint	Football Stadium	Korean Restaurant	Malay Restaurant	Dessert Shop

Discussions

I've noticed that there're more Café shops in Beecroft. There're very few grocery shops in Beecroft. A business owner might be interested in opening a grocery shop there.

Conclusion

In summary, the data analyse of this topic may bring interesting report to the readers. Some information becomes more intuitive to readers. That may help them to understand more of these suburbs.